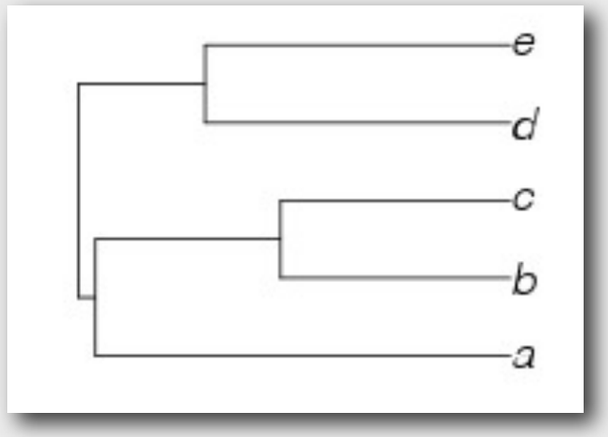
Publishing Re-useable Phylogenetic Trees, in Theory and Practice



```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

Brian O'Meara, Jamie Whitacre, Ross Mounce, Dan Rosauer, Rutger Vos, Arlin Stoltzfus

Tuesday, June 21, 2011

To the extent that we rely so much on symbolically encoded information— which can be stored, endlessly copied, and transmitted worldwide (unlike the case for stone tablets or paper notebooks)— all the world's scientists today could build on all the science done the previous day, or the previous hour. There is an enormous potential for re-use of comparative data and phylogenies.

Why re-use an old tree?

- To evaluate new comparative data (evolution, ecology, biogeography, disease)
- To use as inputs for building larger trees (constraints, supertrees, megatrees)
- To aggregate data in useful resources (TimeTree)
- To study the effects of methodology (priors on tree shape)

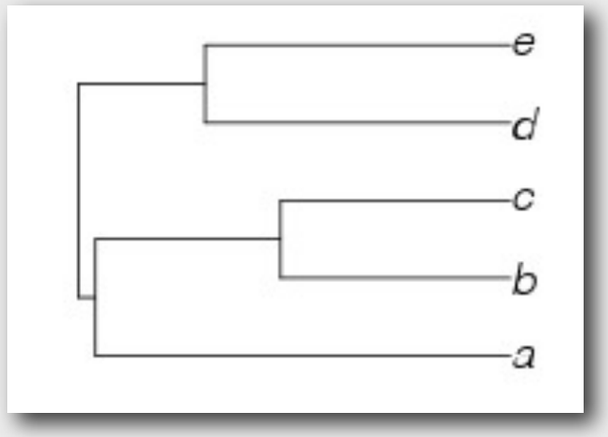
Tuesday, June 21, 2011

There are a number of reasons that users might want to re-use a tree (list).

Just as trees can be re-used, so can other parts of a comparative study, including aligned characters, or phylogenetic results other than trees, such as inferred dates or ancestral character states.

From examining the scientific literature, we know that comparative data and trees get re-used sometimes. However, what we have learned from users suggests that most attempts to discover, access and re-use comparative data and trees end in disappointment.

Publishing Re-useable Phylogenetic Trees, in Theory and Practice



```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

Brian O'Meara, Jamie Whitacre, Ross Mounce, Dan Rosauer, Rutger Vos, Arlin Stoltzfus

Tuesday, June 21, 2011

My co-authors and I are part of a loose network of people— a network in which NESCent plays a major role— interested in facilitating re-use of comparative data and trees. Re-use of scientific information is partly a cultural thing that depends on awareness, values and skills of users, and partly a a technology thing that relies on the tools that support a cycle of transmission, archiving, storage, discovery, access, retrieval and use. We are interested in both the human aspect and the technology aspect.

What we've been doing

- Evaluating capacities and ease-of-use of
  - File formats
  - Archives
  - Interop tools and strategies
- Reviewing applicable policies and standards

Draft report:

http://wiki.tdwg.org/twiki/bin/view/Phylogenetics/LinkingTrees2010

Tuesday, June 21, 2011

We've been doing several things to try to understand the cycle of re-use and how to enhance it (list).

The results of this are available in a draft report.

What we've been doing

- Evaluating capacities and ease-of-use of
  - File formats
  - Archives
  - Interop tools and strategies
- Reviewing applicable policies and standards
- **Assessing user needs and practices**
  - **Publishing and archiving practices**
  - **Examples of re-use of phylogenetic results**
  - **Perceived needs, barriers to re-use**

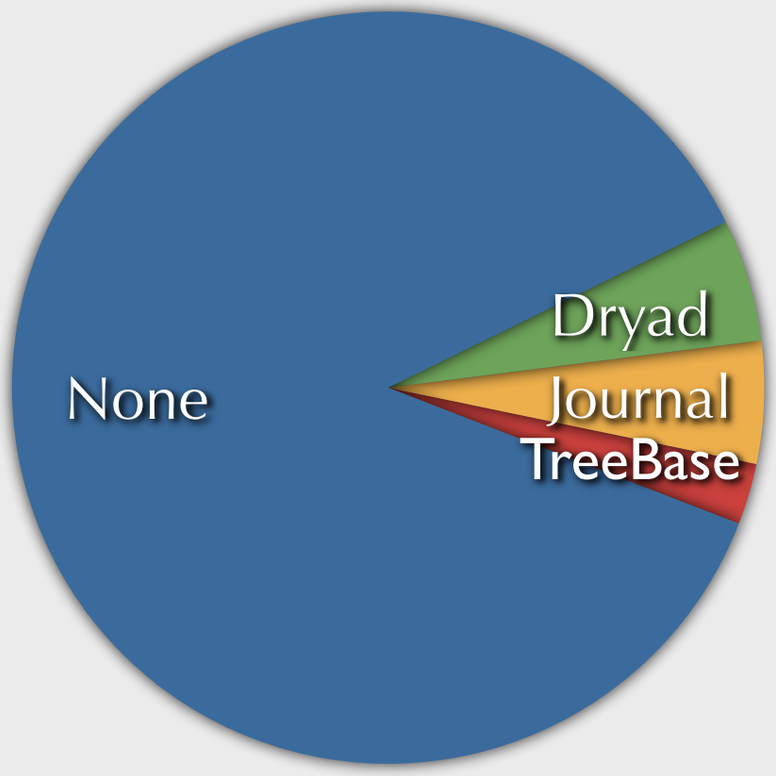
Tuesday, June 21, 2011

What I'm going to talk about today is assessing user needs and practices-- the human aspects of re-use, rather than the technology development aspect. To understand user needs and practices better, we are doing 3 things. Some of us are preparing a survey on barriers to re-use which will be widely distributed later this year. We also have been interviewing users about their experiences. And we have been doing some analysis of the literature. Today I'm going to talk about an analysis of data re-use and archiving done by Brian O'Meara and myself.

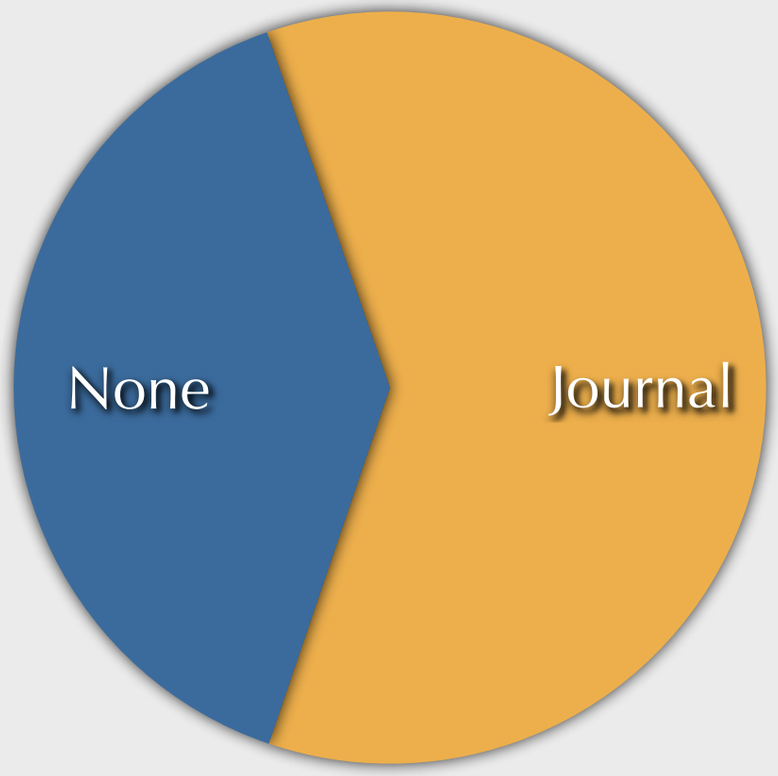
We looked at 40 recent phylogenetics papers, discovered by searching in Web of Science for 2011 papers with "phylogen*" in title or topic. We read each paper and looked for generation, re-use, or archiving of comparative data and trees.

Of 40 recent papers with “phylogeny” in title that created new trees:

Archiving of phylogenies



Archiving of **images** of phylogenies



None Dryad Journal TreeBase

Tuesday, June 21, 2011

Very few papers archive phylogenies in the logical form of a Newick string or some other interoperable format. Instead, authors are often archiving trees as images.

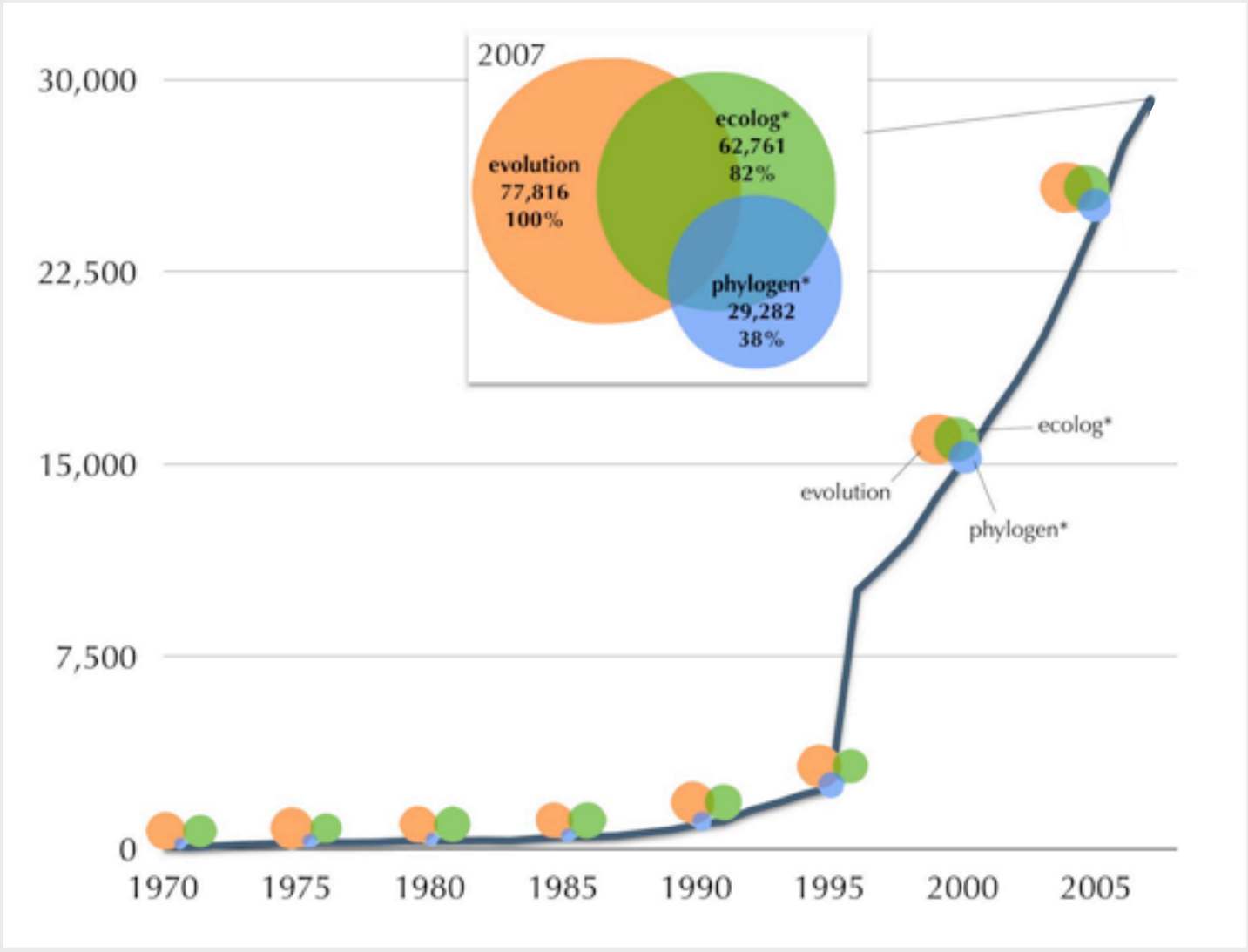
Note: journal SOM + journal figs combined in second figure

Thirsty?

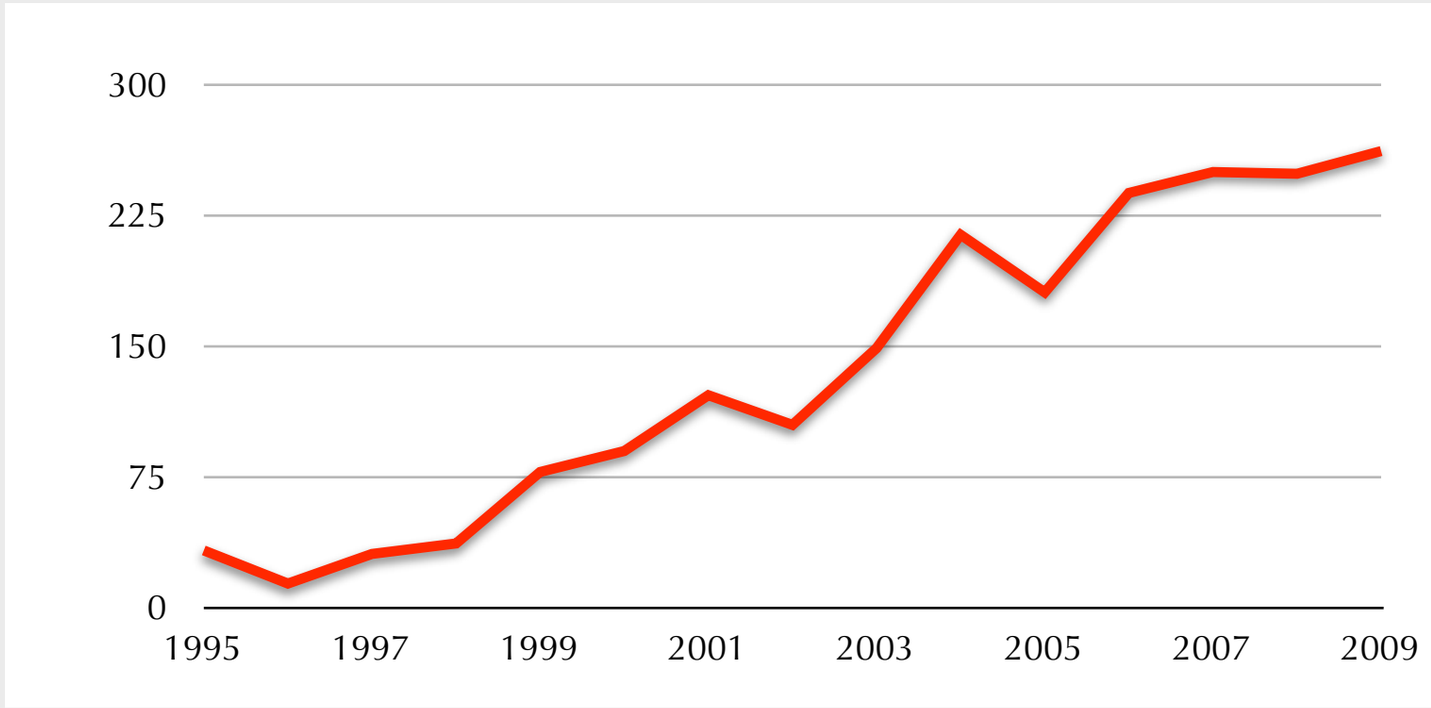
Tuesday, June 21, 2011

Just in case the distinction isn't clear: this is not a glass of water. It is an image of a glass of water.
There is a difference.

Articles with ‘phylogen*’ per year



TreeBase depositions per year

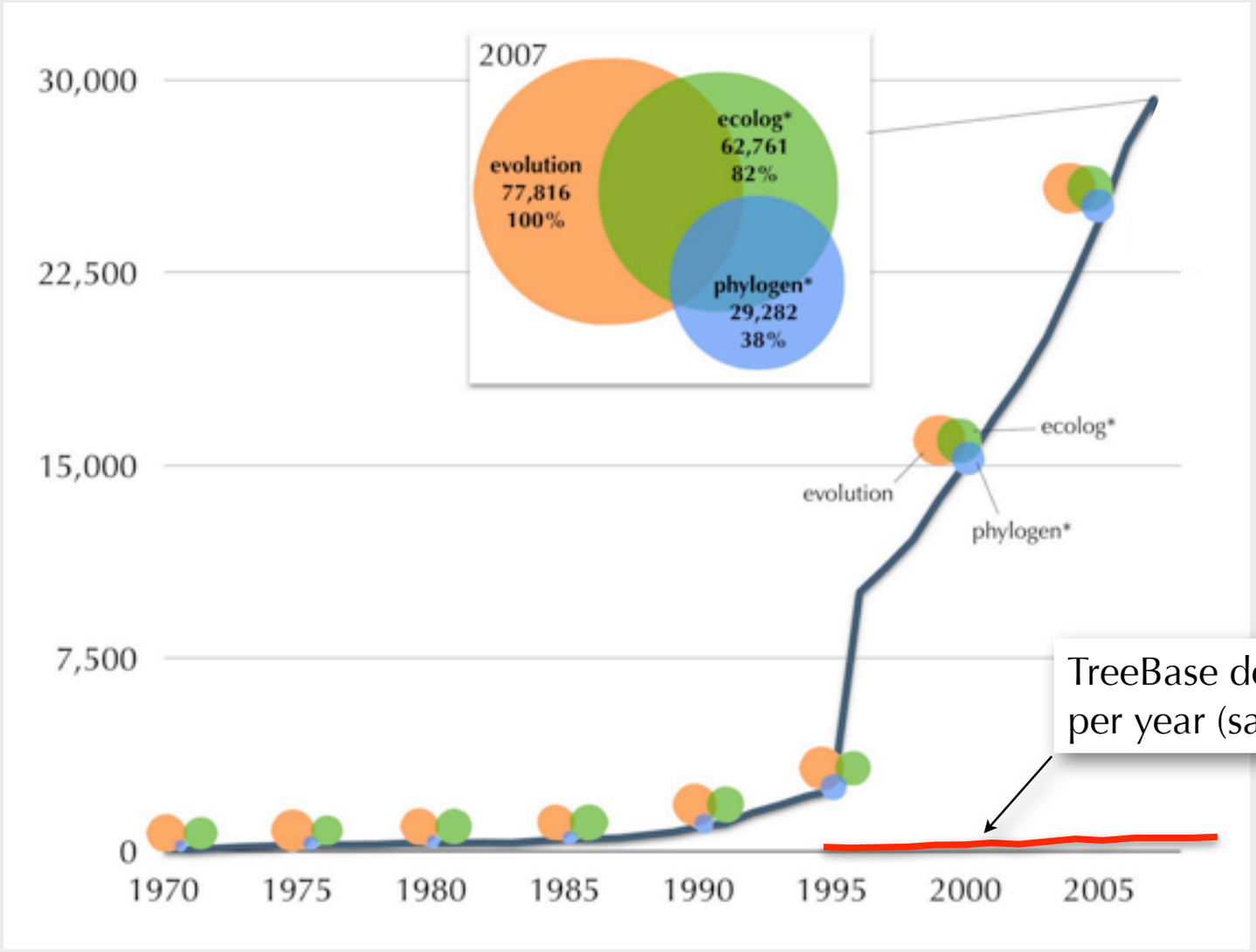


Tuesday, June 21, 2011

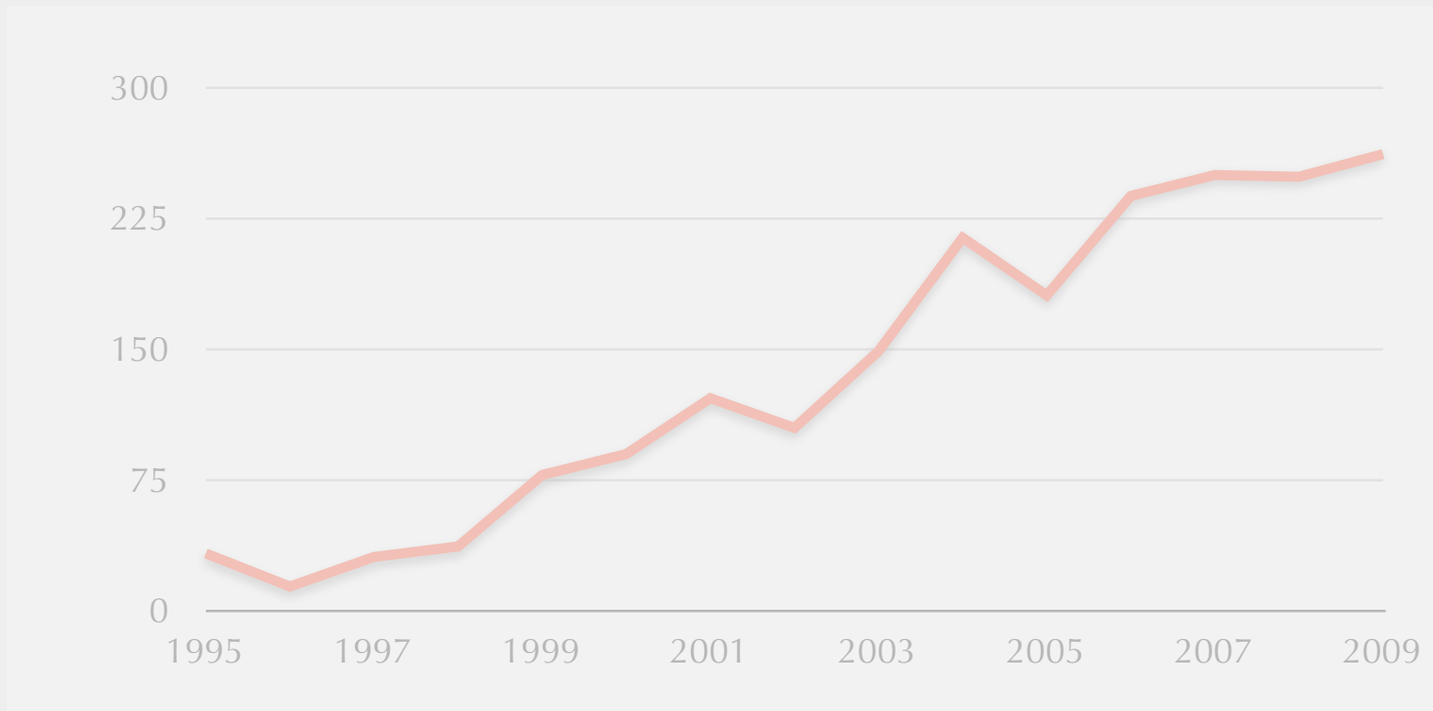
The number of papers referring to phylogeny in title, abstract, etc. has increased enormously. Submissions to TreeBase are growing linearly. I don’t know how many of these papers actually report new phylogenies-- maybe it is 80 % or 40 % or only 20 %.

[References with "phylogen*" in any field in Scopus through time; Venn diagrams showing number of references with "evolution" or "ecolog*" as well as all combinations of these three terms; percent is simply number relative to “evolution”.]

Articles with ‘phylogen*’ per year



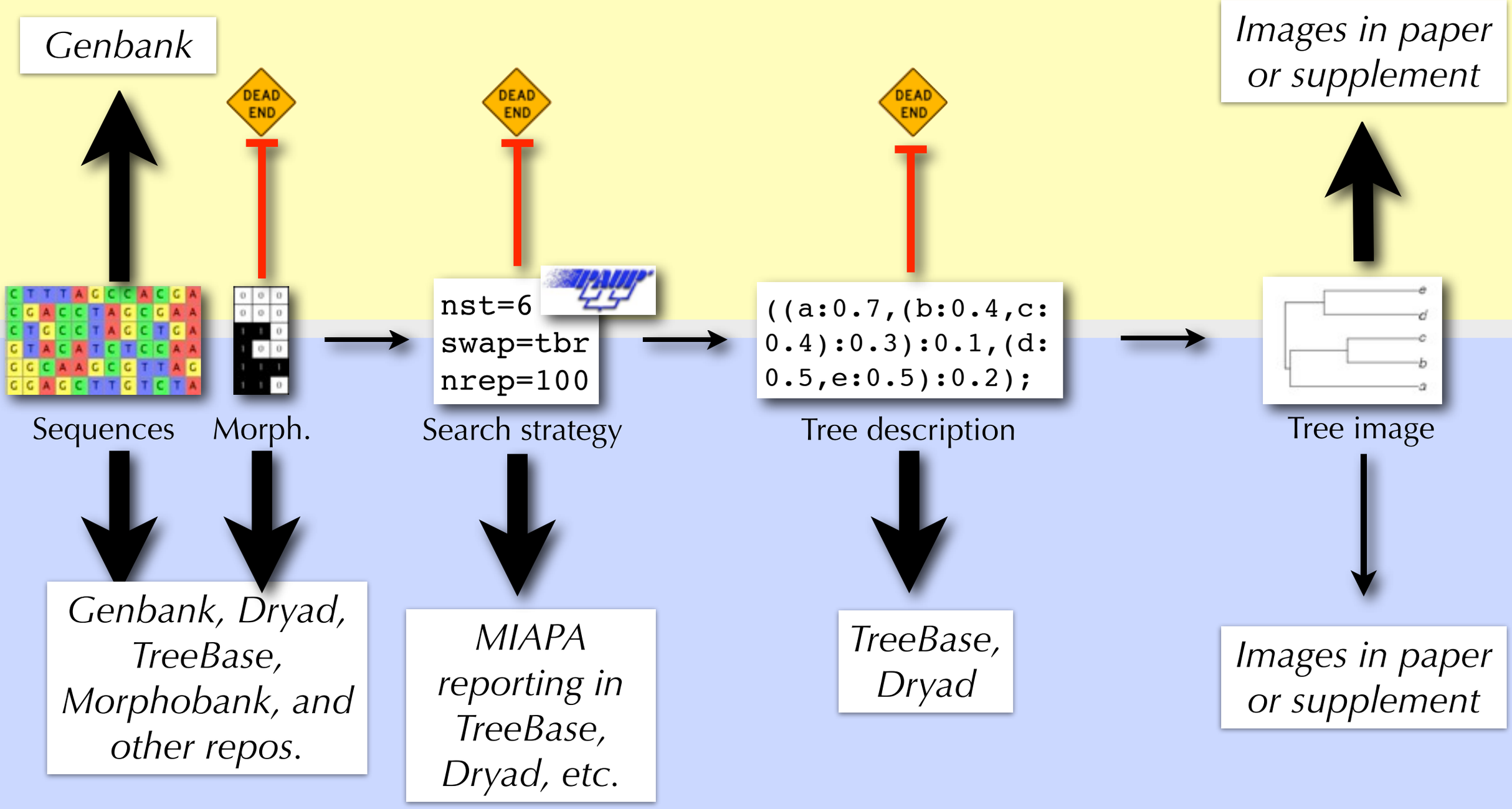
TreeBase depositions per year



Tuesday, June 21, 2011

But even if the number of these papers that report trees is only 1000 per year, it doesn’t seem we are capturing many of them in TreeBase.

Current practice (typical)



Future practice?

Tuesday, June 21, 2011

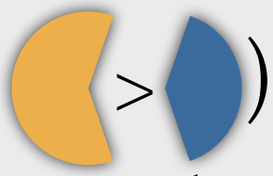
We didn't just look at phylogenies. We also looked at other kinds of information. Currently, we do a good job with deposition of sequence data in GenBank. Not so good with unaligned characters, aligned characters, or usable trees.

Re-use hard	Re-use easy
D. mel.	*Drosophila melanogaster* (even better: *Drosophila melanogaster* Meigen 1830)
Tree image (jpeg, gif, PDF)	Tree description (newick, Nexus, NeXML)
Table image (PDF)	Table text (tab-delimited,.csv, even .xls)
"Available from the author upon request"	"Available from Dryad Digital Repository doi:10.5061/dryad.34984"

Tuesday, June 21, 2011

(list) All the options on the left are more common. The options on the right make data re-use easier. This conclusion isn't based on any carefully controlled research to determined whether full names or highly abbreviated names facilitate re-use. It is just common sense. Anyone who tries to get data from PDF tables quickly learns that copying and pasting from a PDF table is easy, but strange things sometimes happen, and it is difficult to validate whether the contents have been transmitted correctly. A text table, csv, or even a spreadsheet is more interoperable.

Signs of hope

- Many () include pictures of trees in supplementary material. So, desire to save trees, just wrong format.
- NSF data management plans
- [Near] universal GenBank deposition
- Journal requirements
- Dryad (incl. journal support)
- Ongoing work on MIAPA
- TreeBase 2.0 and competitors (i.e., Phylografter)

Tuesday, June 21, 2011

Now, there are also signs of hope (list).

User stories

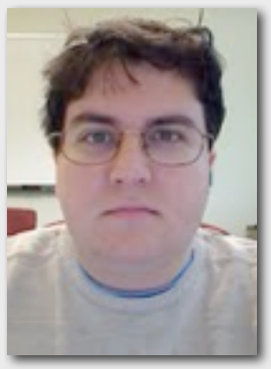
http://www.evoio.org/wiki/BarriersToReUse

- users sometimes go to extraordinary lengths
  - to replicate an analysis before re-using a tree
  - to re-encode high-value aligned character data
  - name ambiguity or mismatch (alignment vs. tree) in preserved records is a major issue
  - software exists to recode trees from images (TreeSnatcher)

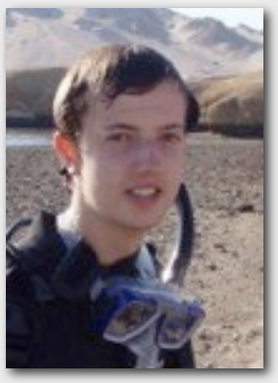
Tuesday, June 21, 2011

The previous few slides have focused on results of the literature analysis that Brian and I did.

I would like to call attention briefly to some of the work of an overlapping group, the MIAPA survey group, which collected user stories in preparation for developing a survey on barriers to re-use. We learned some fascinating things from talking to scientific users directly about their experiences with data re-use (list).


Brian


Jamie


Ross


Dan


Rutger


Arlin

Thanks

- Bill Piel
- **MIAPA survey group** (Sudhir Kumar, Ross Mounce, Rutger Voss, Emily Gillespie, Nico Cellinese, Enrico Pontelli, Arlin Stoltzfus)
- Biodiversity Information Standards (TDWG) and the **TDWG Phylogenetic Standards Interest Group**

Tuesday, June 21, 2011

Finally,
thanks.

Getting involved

- Groups interested in standards:
  - Phylogenetic Standards Interest Group of TDWG
  - MIAPA (miapa-discuss@googlegroups.com)
- Open-source projects relevant to phylo-interop
  - NeXML
  - TreeBase
  - Dryad
  - CDAO
  - others
- Upcoming initiatives
  - MIAPA survey on barriers to re-use of phylogenetic results
  - NESCent's HIP (hackathons+interop+phylo) working group